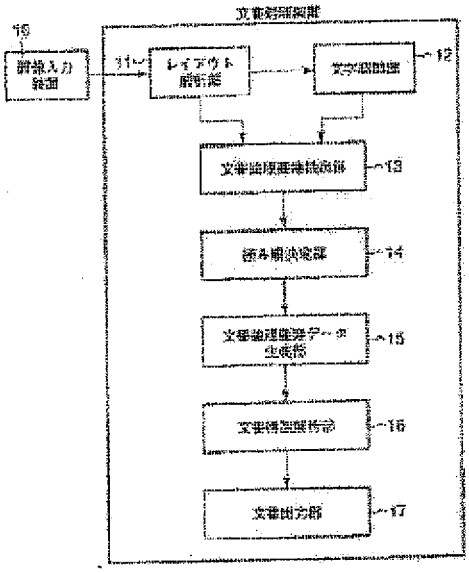
DOCUMENT PROCESSOR AND DOCUMENT PROCESSING METHOD

Publication number: JP2003288334 (A)
Publication date: 2003-10-10
Inventor(s): ISHITANI YASUTO
Applicant(s): TOKYO SHIBAURA ELECTRIC CO
Classification:
- international: ***G06F17/21; G06K9/20;*** **G06F17/21; G06K9/20;** (IPC1-7): G06F17/21; G06K9/20
- European:
Application number: JP20020093092 20020328
Priority number(s): JP20020093092 20020328

Abstract of **JP 2003288334 (A)**

PROBLEM TO BE SOLVED: To generate a structured document such as an XML (extensible markup language) document and an HTML (hypertext markup language) document, by setting in appropriate places document logic elements other than sentence such as graphs, and tables, contained in a printing document consisting of a plurality of pages. ; SOLUTION: The device extracts a paragraph area and a graph area by analyzing document graphs in layout corresponding to a printing document with a layout analyzing part 11 while segmenting characters in the paragraph area to recognize and process with a character recognizing part 12. It extracts a document logic element area from the paragraph area by providing a character recognizing result and a layout analyzing result to a document logic element extracting part 13, and carries out order setting respectively to a document logic element area and a graph/table area with a reading order setting part 14. Then, it extracts a document structure by grouping respectively the document logic element area and the graph/table area with a document structure analyzing part 16, and generates the structure document by changing the appearance position of an area corresponding to the document logic elements other than sentence in the document structure and providing to a document output part 17. ; COPYRIGHT: (C)2004,JPO



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－288334
（P2003－288334A）
(43)公開日　平成15年10月10日(2003.10.10)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/21	５３０	Ｇ０６Ｆ　17/21	５３０Ａ	５Ｂ００９
	５０１		５０１Ｚ	５Ｂ０２９
Ｇ０６Ｋ　9/20	３４０	Ｇ０６Ｋ　9/20	３４０Ｌ	

審査請求　有　　請求項の数10　ＯＬ　（全 13 頁）

(21)出願番号　特願2002－93092(P2002－93092)

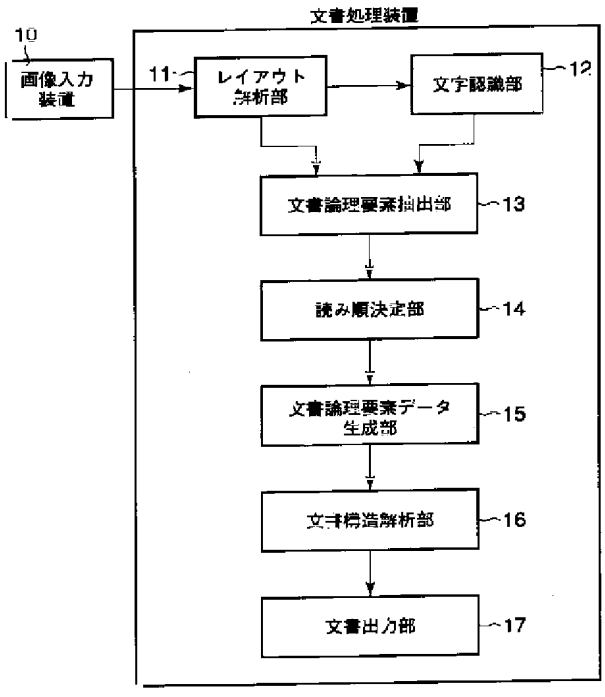
(22)出願日　平成14年３月28日(2002.3.28)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(72)発明者　石谷　康人
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　100058479
弁理士　鈴江　武彦　（外６名）

Ｆターム(参考)　5B009 QA06 QB14 TA11
5B029 CC28 CC29 EE08

(54)【発明の名称】　文書処理装置及び文書処理方法

(57)【要約】
【課題】複数ページからなる印刷文書に含まれる図、表などの浮遊要素を適切な位置に設定してＸＭＬ文書やＨＴＭＬ文書などの構造化文書を生成する。
【解決手段】印刷文書に対応した文書画像をレイアウト解析部１１にてレイアウト解析して文章領域と図表領域を抽出すると共に文字認識部１２にて文章領域中の文字を切り出して認識処理する。文字認識結果とレイアウト解析結果を文書論理要素抽出部１３に与えて文章領域から文書論理要素領域を抽出し、読み順設定部１４にて文書論理要素領域と図表領域の各領域に対して順序付けを行う。そして、文書構造解析部１６にて文書論理要素領域と図表領域の各領域をグループ化して文書構造を抽出し、その文書構造の中で浮遊要素に相当する領域の出現位置を変更して文書出力部１７に与えて構造化文書を生成する。



【特許請求の範囲】
【請求項１】　複数ページからなる印刷文書を処理対象とし、その印刷文書に対応した文書画像から所定形式の構造化文書を生成する文書処理装置であって、
上記文書画像をレイアウト解析して文章領域と図表領域に相当する部分画像を抽出するレイアウト解析手段と、
このレイアウト解析手段によって得られた文章領域の部分画像から文字を切り出して認識処理する文字認識手段と、
この文字認識手段の文字認識結果と上記レイアウト解析手段のレイアウト解析結果に基づいて、上記文章領域から文書見出し、章節見出し、段落、箇条書き、ヘッダ、フッタ、図表キャプション、脚注、数式などの文書論理要素領域を抽出する文書論理要素抽出手段と、
この文書論理要素抽出手段によって得られた文書論理要素領域と上記図表領域の各領域に対して順序付けを行う読み順設定手段と、
この読み順設定手段によって設定された各領域の順序と領域種別、領域間の配置関係に基づいて上記文書論理要素領域と上記図表領域の各領域をグループ化することにより文書構造を抽出する文書構造抽出手段と、
この文書構造抽出手段によって得られた文書構造の中でグループ化された領域あるいは単独の領域の出現位置を変更する文書構造変更手段と、
この文書構造変更手段によって変更された文書構造に基づいて所定形式の構造化文書を生成する構造化文書生成手段とを具備したことを特徴とする文書処理装置。
【請求項２】　上記文書構造変更手段は、変更すべき領域の位置をオペレータに指定させることを特徴とする請求項１記載の文書処理装置。
【請求項３】　上記文書論理要素抽出手段によって抽出された各種文書論理要素の中で特定の文書論理要素を不要要素として除去する不要要素除去手段を備えたことを特徴とする請求項１記載の文書処理装置。
【請求項４】　上記文書構造変更手段によって領域の位置が変更された際に、その領域の前後に存在する同一属性の文書論理要素を統合する要素統合手段を備えたことを特徴とする請求項１記載の文書処理装置。
【請求項５】　複数ページからなる印刷文書を処理対象とし、その印刷文書に対応した文書画像から所定形式の構造化文書を生成する文書処理装置であって、
上記文書画像をレイアウト解析して文章領域と図表領域に相当する部分画像を抽出するレイアウト解析手段と、
このレイアウト解析手段によって得られた文章領域の部分画像から文字を切り出して認識処理する文字認識手段と、
この文字認識手段の文字認識結果と上記レイアウト解析手段のレイアウト解析結果に基づいて、上記文章領域から文書見出し、章節見出し、段落、箇条書き、ヘッダ、フッタ、図表キャプション、脚注、数式などの文書論理要素領域を抽出する文書論理要素抽出手段と、
この文書論理要素抽出手段によって得られた文書論理要素領域と上記図表領域の各領域に対して順序付けを行う読み順設定手段と、
この読み順設定手段によって設定された各領域の順序と領域種別、領域間の配置関係に基づいて上記文書論理要素領域と上記図表領域の各領域をグループ化することにより文書構造を抽出する文書構造抽出手段と、
この文書構造抽出手段によって得られた文書構造の中で上記図表領域に関するグループの出現位置を変更する文書構造変更手段と、
この文書構造変更手段によって変更された文書構造に基づいて所定形式の構造化文書を生成する構造化文書生成手段とを具備したことを特徴とする文書処理装置。
【請求項６】　上記文書構造変更手段は、上記図表領域に関するグループを当該図表が含まれる章節の末尾に配置することを特徴とする請求項５記載の文書処理装置。
【請求項７】　上記文書構造変更手段は、上記図表領域に関するグループを文書の末尾に配置することを特徴とする請求項５記載の文書処理装置。
【請求項８】　上記文書構造変更手段は、上記図表領域に関するグループを当該図表を最初に参照している文書論理要素の直後に配置することを特徴とする請求項５記載の文書処理装置。
【請求項９】　上記文書構造変更手段は、上記図表領域に関するグループを当該図表を最も多く参照している文書論理要素の直後に配置することを特徴とする請求項５記載の文書処理装置。
【請求項１０】　複数ページからなる印刷文書を処理対象とし、その印刷文書に対応した文書画像から所定形式の構造化文書を生成する文書処理方法であって、
上記文書画像をレイアウト解析して文章領域と図表領域に相当する部分画像を抽出し、
このレイアウト解析結果として得られた文章領域の部分画像から文字を切り出して認識処理し、
この文字認識結果と上記レイアウト解析結果に基づいて、上記文章領域から文書見出し、章節見出し、段落、箇条書き、ヘッダ、フッタ、図表キャプション、脚注、数式などの文書論理要素領域を抽出し、
上記文書論理要素領域と上記図表領域の各領域に対して順序付けを行い、
各領域の順序と領域種別、領域間の配置関係に基づいて上記文書論理要素領域と上記図表領域の各領域をグループ化することにより文書構造を抽出し、
上記文書構造の中でグループ化された領域あるいは単独の領域の出現位置を変更し、
この変更後の文書構造に基づいて所定形式の構造化文書を生成することを特徴とする文書処理方法。
【発明の詳細な説明】
【０００１】

【発明の属する技術分野】本発明は、オフィスや家庭などで流通している新聞記事、雑誌、科学技術文献、書籍、オフィス文書、公文書などの印刷文書を処理の対象としており、この種の印刷文書に記載されている内容を抽出・構造化する文書処理装置に係り、特に文章の他に図面や表などが混在した印刷文書を例えばＸＭＬ（extensible markup language）文書やＨＴＭＬ（hypertext markup language）文書などの構造化文書に変換する場合に用いて好適な文書処理装置及び文書処理方法に関する。

【０００２】

【従来の技術】オフィスや家庭などでは、新聞記事、雑誌、科学技術文献、書籍、オフィス文書、公文書など、様々な印刷文書が流通している。この種の印刷文書をスキャナ等の入力装置から文書画像（イメージ）としてコンピュータに取り込み、その画像情報を解析して文字領域から文字パターンを切り出し、文字認識技術により各文字をコード化することにより、印刷文書の内容を電子化・再利用したいという要求がある。この場合、従来の技術では、文字認識の対象となる文章領域のみに着目して文書の電子化が行われていた。このような技術の例はいくつかあるが、代表的なものを挙げて説明する。

【０００３】ここで、文献「黄瀬他：“文書画像構造解析のための知識ベースの一構成法”、情処学論、Vol.34，No.1，PP75-87，(1993-1)」によれば、文書構造とは、“レイアウト構造”と“論理構造”から構成されており、“レイアウト構造”とはテキスト、図、写真、表などの文書要素の幾何情報（位置情報と大きさ情報）に関する階層構造のことであり、ブロック領域などのレイアウトオブジェクトを要素として持つと定義される。“論理構造”とは論理意味情報（コンテンツ）に関する階層構造のことであり、章節などの論理オブジェクトを要素として持つと定義される。このような定義を念頭において、以下、いくつかの従来技術に触れてみる。

【０００４】（１）「S.Tsujimoto: ”Major Components of a Complete Text Reading System,”Proceedings of THE IEEE, Vol.80, No.7, July, 1992」

この論文では、レイアウト解析によって得られたレイアウトオブジェクトの幾何的階層構造に対して、いくつかの一般的なルールを適用することにより、論理構造に変換する方式について説明している。この場合、論理構造は木構造で表現されるが、それをルートから辿っていくと読み順が得られる。

【０００５】（２）「駱他：“ルールベースの適用による日本語新聞紙紙面の構造認識”、信学論D-II，Vol.J75-D-II，No.9，pp.1514-1525，(1992-9)」

これは、日本語新聞のレイアウトオブジェクトを隣接関係グラフで表現し、ルールに基づいてこのグラフを解釈することでタイトル（見出し）、本文、セパレータ、写真、図表で構成される記事を個別に抽出するというものである。

【０００６】（３）「山下他：“モデルに基づいた文書画像のレイアウト理解”、信学論D-II，Vol.J75-D-II，No.10，pp.1673-1681，(1992-10)」

これは、レイアウトオブジェクトと１対１に対応した論理オブジェクトについて、表形式で簡単に表現されたモデルを入力文書のレイアウト解析結果に適用して、簡単な論理構造を抽出するものである。

【０００７】（４）「黄瀬他：“文書画像構造解析のための知識ベースの一構成法”、情処学論、Vol.34，No.1，PP75-87，(1993-1)」

これは、レイアウト構造と論理構造とその対応関係を表す文書モデルを用いて入力文書に対して推論を適用することにより文書構造を抽出する。文書モデルは、構造の階層性を記述できるフレーム表現を採用しており、センタリングなどのレイアウト記述を可能とし、各構成要素の変動の記述も可能にしている。

【０００８】（５）「山田：“文書画像のＯＤＡ論理構造化文書への変換方式”、信学論D-II，Vol.J76-D-II，No.11，pp.2274-2284，(1993-11)」

これは、入力文書をＯＤＡ機能標準PM（プロセッサブルモード）26文書に自動マッピングする方式である。節構造解析により、複数ページから多段の章・節・段落を抽出・構造化し、表示属性解析により、字下げ、揃え、ハードリターン、オフセットを抽出する。また、ヘッダ／フッタ解析により、文書クラスの同定も可能としている。

【０００９】（６）「建石：“確率文法を用いた文書論理構造の解釈法”、信学論D-II，Vol.J79-D-II，No.5，pp.687-697，(1996-5)」

これは、確率文法の枠組を用いて、複数ページに渡る章節構造とリスト構造を抽出するものである。

【００１０】

【発明が解決しようとする課題】従来技術では、単ページに相当する１つの文書画像に対してのみ文書構造解析処理を適用していることがあった。また、複数ページ分の文書画像を対象としている場合であっても、図、表、写真、脚注といったような文書論理構造（章節構造やリスト（箇条書き）構造）には規定されない浮遊要素の取り扱いを無視していたり、あるいは、その浮遊要素を不適切に取り扱ったりしていた。近年、インターネットの普及に伴い、印刷文書に記載されている内容をＸＭＬ文書やＨＴＭＬ文書などのタグ付きされた構造化文書に変換する要求が高まっており、その際に、印刷文書中に含まれる図、表、写真、脚注などの浮遊要素（文章とは別の文書論理要素）をオペレータが意図する位置に適切に取り込んだ形で構造化することが望まれる。

【００１１】そこで、本発明の目的は、複数ページからなる印刷文書に含まれる文章とは別の図、表などの浮遊要素を適切な位置に設定してＸＭＬ文書やＨＴＭＬ文書

などの構造化文書を生成可能な文書処理装置及び文書処理方法を提供することにある。

【００１２】

【課題を解決するための手段】上記目的を達成するため、本発明は、複数ページからなる印刷文書に対応した文書画像をレイアウト解析して文章領域と図表領域に相当する部分画像を抽出するレイアウト解析手段と、このレイアウト解析手段によって得られた文章領域の部分画像から文字を切り出して認識処理する文字認識手段と、この文字認識手段の文字認識結果と上記レイアウト解析手段のレイアウト解析結果に基づいて、上記文章領域から文書見出し、章節見出し、段落、箇条書き、ヘッダ、フッタ、図表キャプション、脚注、数式などの文書論理要素領域を抽出する文書論理要素抽出手段と、この文書論理要素抽出手段によって得られた文書論理要素領域と上記図表領域の各領域に対して順序付けを行う読み順設定手段と、この読み順設定手段によって設定された各領域の順序と領域種別、領域間の配置関係に基づいて上記文書論理要素領域と上記図表領域の各領域をグループ化することにより文書構造を抽出する文書構造抽出手段と、この文書構造抽出手段によって得られた文書構造の中でグループ化された領域あるいは単独の領域の出現位置を変更する文書構造変更手段と、この文書構造変更手段によって変更された文書構造に基づいて所定形式の構造化文書を生成する構造化文書生成手段とを具備して構成されるものである。

【００１３】このような構成によれば、文書画像に含まれる文章と図表の各領域をグループ化した文書構造の中で任意の領域の出現位置を変更できるので、例えば図や表といった文章とは別の浮遊要素を適切な位置に設定してＸＭＬ文書やＨＴＭＬ文書などのタグ付けされた構造化文書を生成することができる。この場合、変更すべき領域の位置をオペレータに指定させることで、オペレータの意図する文書構造で構造化文書を得ることができる。

【００１４】また、上記文書論理要素抽出手段によって抽出された各種文書論理要素の中で特定の文書論理要素を不要要素として除去する不要要素除去手段を備えることで、例えばヘッダやフッタを不要要素として予め指定しておけば、これを除去した形で構造化文書を得ることができる。

【００１５】また、上記文書構造変更手段によって領域の位置が変更された際に、その領域の前後に存在する同一属性の文書論理要素を統合する要素統合手段を備えることで、例えば文章と文章の間に挿入されていた図表などの浮遊要素を他の場所に移動させた際に、その図表の前後にあった２つの文章を１つの連続した文章として統合した形で構造化文書を得ることができる。

【００１６】また、本発明は、複数ページからなる印刷文書を処理対象とし、その印刷文書に対応した文書画像から所定形式の構造化文書を生成する文書処理装置であって、上記文書画像をレイアウト解析して文章領域と図表領域に相当する部分画像を抽出するレイアウト解析手段と、このレイアウト解析手段によって得られた文章領域の部分画像から文字を切り出して認識処理する文字認識手段と、この文字認識手段の文字認識結果と上記レイアウト解析手段のレイアウト解析結果に基づいて、上記文章領域から文書見出し、章節見出し、段落、箇条書き、ヘッダ、フッタ、図表キャプション、脚注、数式などの文書論理要素領域を抽出する文書論理要素抽出手段と、この文書論理要素抽出手段によって得られた文書論理要素領域と上記図表領域の各領域に対して順序付けを行う読み順設定手段と、この読み順設定手段によって設定された各領域の順序と領域種別、領域間の配置関係に基づいて上記文書論理要素領域と上記図表領域の各領域をグループ化することにより文書構造を抽出する文書構造抽出手段と、この文書構造抽出手段によって得られた文書構造の中で上記図表領域に関するグループの出現位置を変更する文書構造変更手段と、この文書構造変更手段によって変更された文書構造に基づいて所定形式の構造化文書を生成する構造化文書生成手段とを具備して構成されるものである。

【００１７】このような構成によれば、文書画像に含まれる文章と図表の各領域をグループ化した文書構造の中で図表、図表キャプション、図表脚注といったような図表領域に関する各要素を１つのグループとして、その出現位置を変更してＸＭＬ文書やＨＴＭＬ文書などのタグ付けされた構造化文書を生成することができる。この場合、図表領域の変更位置として、当該図表が含まれる章節の末尾や文書の末尾に配置したり、当該図表を最初に参照している文書論理要素の直後に配置したり、当該図表を最も多く参照している文書論理要素の直後に配置することができる。

【００１８】

【発明の実施の形態】以下、図面を参照して本発明の一実施形態を説明する。

【００１９】本発明の文書処理装置では、新聞記事、雑誌、科学技術文献、書籍、オフィス文書、公文書などの複数ページからなる印刷文書を処理対象とし、図1に示すように、この種の印刷文書をスキャナ等にて読み取った文書画像（文書イメージ）を元にＸＭＬ文書やＨＴＭＬ文書などの所定形式の構造化文書を生成するものである。ＨＴＭＬは、ハイパーテキストを記述するための言語であって、タグ（＜＞で囲んだ予約語）を用いて、文書の構造、画像などのファイルのある場所などを記述する。ＸＭＬは、ＨＴＭＬの後継言語であり、ＳＧＭＬ（standard generalized markup language）の持つ拡張機能をＷｅｂ上でも利用可能とした言語である。ＨＴＭＬとの違いは、ユーザが独自のタグを使ってデータの属性情報や論理構造を定義することができる。

【００２０】図１（ａ）は本装置の処理対象となる印刷文書に対応した文書画像の一例を示す図であり、ここでは３ページ分の文書画像が示されている。この文書画像には、文章（テキスト）の他に、図や表などが含まれている。また、同図（ｂ）はその文書画像に対応した構造化文書の一例を示す図であり、タグ付けされたＸＭＬ文書やＨＴＭＬ文書などである。

【００２１】図２は本発明の一実施形態に係る文書処理装置の機能構成を示すブロック図である。まず、複数ページ（ここでは３ページ）で構成される印刷文書がスキャナ等からなる画像入力装置１０から連続する複数枚の文書画像として順次入力される。なお、各画像はスキャナ等で事前に２値化処理が行われて２値画像に変換されていても良いし、公知の技術により２値画像に変換されても良い。さらには、例えば特開平５－１７４１８３号公報の「文書画像傾き検出方式」に記載されている技術により、画像の傾きが検出・補正されることにより真直ぐな２値画像に変換されていても良い。以下では、傾きが補正された２値画像を入力画像と呼ぶことにする。印刷文書に対応した複数枚の文書画像は、順次、本発明の文書処理装置に入力されるものとする。

【００２２】図２に示すように、本実施形態における文書処理装置は、レイアウト解析部１１、文字認識部１２、文書論理要素抽出部１３、読み順設定部１４、文書論理要素データ生成部１５、文書構造解析部１６、文書出力部１７から構成される。なお、本装置は、例えば磁気ディスク等の記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されるコンピュータによって実現されるものであり、図２ではコンピュータに搭載されたＣＰＵがプログラムに記述された手順に従って実行する各処理を機能的に示したものである。

【００２３】画像入力装置１０から入力された文書画像は本装置のレイアウト解析部１１に与えられ、その入力画像から図３に示すように文章領域、図領域、表領域などの性質の異なる部分領域がレイアウト要素群として抽出される。図３（ａ）は印刷文書に対応した文書画像、同図（ｂ）はレイアウト解析処理によって得られる上記文書画像の部分領域（レイアウト要素群）を示す図である。

【００２４】部分領域は、図４に示すようにその領域に外接する矩形によって表現されるものとする。この場合、図５に示すように、その外接矩形の位置は左上端の位置座標（ｘ１，ｙ１)と右下端の位置座標（ｘ２，ｙ２)によって表現することができる。レイアウト解析では、縦書きと横書きの文章領域は異なる領域として分離されて出力される。また、１つの文章領域はカラムを跨って抽出されることがないものとする。ただし、段組（パラグラフ）、箇条書き（リスト）、数式、章節タイトルなどの文書論理要素に相当する領域が抽出されてなくても良いものとする。

【００２５】ここで、文章領域では、図４に示すように、文字行領域が順序付けられて抽出され、また、各文字行領域では文字領域が同様に順序付けられて抽出される。この文字行領域と文字領域はそれぞれを外接する矩形により表現される。図中のＴＢは文章領域、Ｓｔｒは文字行領域、Ｃｈは文字領域を示している。図４の例では、文章領域ＴＢに５つの文字行領域Ｓｔｒ１～Ｓｔｒ５が存在し、その中の文字行領域Ｓｔｒ１には５つの文字領域Ｃｈ１～Ｃｈ５が存在している。

【００２６】ここで、文章領域と文字行領域と文字領域がそれぞれ階層的に、例えば図６のような木構造により記述される。ここで、レイアウト解析部１１は、例えば特開平９－１６７２３３号公報の「画像処理方法および画像処理装置」に記載されているような構成であっても良い。この場合、レイアウト解析部１１で文字認識処理が行われて、文章領域の各文字がコード化されるようになっていても良いし、レイアウト解析の直後に文字認識処理が実施されて文章領域の各文字がコード化されるようになっていても良い。

【００２７】文字認識部１２は、例えば「有吉：“動的な仮説生成・検証による日本語印刷文書からの文字の切り出し”，電子情報通信学会技術報告，PRU93-47，pp.33-40，1993.」により実現されていても良い。この場合、レイアウト解析で得られた文字行領域から個々の文字領域を切り出した後に、文字領域内の文字パターンを文字コードに変換する。このとき文字認識結果は最終的に読み順に並んだ状態で出力される。

【００２８】次に、レイアウト解析部１１によって得られたレイアウト解析結果が文字認識結果と共に文書論理要素抽出部１３に与えられる。文書論理要素抽出部１３では、図７に示すように、文章領域から文書見出し、段落、ヘッダ、フッタ、キャプション（図表標題）、箇条書き、章節見出し、脚注、数式などの文書論理要素領域を抽出する。

【００２９】図７（ａ）は文書画像の部分領域（レイアウト要素群）、同図（ｂ）はその部分領域から抽出された文書論理要素領域を示す図であり、例えば１ページ目の文書における１段目の「文章領域」から「ヘッダ領域」、２段目の「文章領域」から「文書見出し領域」、３段目の「文章領域」から「章節見出し領域」、「段組領域」、「章節見出し領域」、「段組領域」、「箇条書き領域」、「箇条書き領域」、３段目の「文章領域」から「フッタ領域」が抽出された状態が示されている。この場合、図１（ａ）に示すように、１段目の「文章領域」から抽出された「ヘッダ領域」には「論文Ｎｏ．１　Ｖｏｌ．１」、２段目の「文章領域」から抽出された「文書見出し領域」には「文書処理方法」、３段目の「文章領域」から抽出された「章節見出し領域」には「１．はじめに」、４段目の「文章領域」から抽出され

た「フッタ領域」には「１」といった文字列が記述されており、その文字認識結果は文字認識部１２から与えられている。

【００３０】なお、文書論理要素抽出方法としては、例えば特開平１１－２５００４１号公報の「文書処理装置および文書処理方法」に記載されている方法が用いられるものとし、レイアウト解析結果と文字認識結果に基づいて上述した文書論理要素領域が出力される。この場合、各文書論理要素の領域は外接矩形で表現され、図４乃至図６に示したように、その外接矩形の内部に文字行領域が存在し、文字行領域の内部に文字領域が存在し、文書論理要素領域と文字行領域と文字領域が木構造により階層的に記述されている。そして、各文書論理要素領域には、文書見出し、段落、ヘッダ、フッタ、キャプション（図表標題）、箇条書き、章節見出し、脚注、数式といった種類情報が論理属性として割り当てられる。

【００３１】また、表領域については、例えばＯｉといった表領域があり、その外接矩形の左上端位置座標を（ＯｉＸ１，ＯｉＹ１）、右下端位置座標を（ＯｉＸ２，ＯｉＹ２）とし、ｔｈ３およびｔｈ４を所定の閾値とした場合に、以下のような条件の満たす文章領域を表キャプションとして抽出する。

（１）左上端位置座標（（ＯｉＸ１－ｔｈ３），（ＯｉＹ１－ｔｈ４））と右下端位置座標（ＯｉＸ１，ＯｉＹ１）で形成される表領域の上部に位置する探索矩形領域に含まれている。

（２）文字認識結果に「表１」といったような特定の文字が含まれている。

【００３２】そして、このようして抽出した表キャプションと表領域との間に位置する文章領域や、左上端位置座標（ＯｉＸ２，ＯｉＹ２）および右下端位置座標（ＯｉＸ２＋ｔｈ３，ＯｉＹ２＋ｔｈ４）で形成される表領域の下部に位置する探索矩形領域に含まれる文章領域を表脚注として抽出する。表キャプションを抽出するための探索矩形領域は前述のように表の上部にあっても良いし、表の下部にあっても良い。図キャプションと図脚注についても、上記同様のアプローチで抽出することができる。

【００３３】次に、文書論理要素抽出部１３によって抽出された文書論理要素領域がレイアウト解析部１１で得られたレイアウト要素領域と共に読み順設定部１４に入力されて、レイアウト要素領域と文書論理要素領域の配置関係や幾何情報に基づいて文書論理要素領域に対する順序付けがなされる。この読み順設定方法としては、例えば特開平１１－２５００４１号公報の「文書処理装置および文書処理方法」に記載されている方法が用いられる。その結果、例えば図８（ａ）のように、各文書毎に上から順に各領域に対して読み順が付される。

【００３４】このようにして、複数枚の文書画像のそれぞれに対し、レイアウト解析処理、文字認識処理、文書論理要素抽出処理、読み順設定処理が施されると、図９に示すように、文書論理要素データ生成部１５により各領域毎に論理属性、読み順、文字認識結果などからなる文書論理要素データが生成されて文書構造解析部１６に供給される。図９（ａ）は文書画像の文書論理要素領域とその読み順を示す図、同図（ｂ）はその文書論理要素領域の木構造を示す図、同図（ｃ）はその文書論理要素領域のデータ構成を示す図である。

【００３５】文書構造解析部１６には、例えば全ての文書画像のレイアウト解析から読み順設定までの処理がなされた後に、各文書画像に対応した文書論理要素データが一括して入力されるようになっていても良い。この場合、文書毎に付与された読み順情報は、図８（ｂ）に示すようにページの連続性に伴って最初のページの先頭文書論理要素から最終ページの最後の文書論理要素まで連続した読み順が付与されるように変更される。

【００３６】文書構造解析部１６は、図１０に示すように、不要文書論理要素除去部２１、文書論理要素グループ化部２２、文書構造抽出部２３、文書構造変更部２４、文書論理要素統合部２５から構成される。

【００３７】まず、不要文書論理要素除去部２１により、例えばヘッダやフッタのように、全てのページにおいて重複して出現し、複数ページを統合して１つの構造化文書を生成する場合に不要となる文書論理要素を除去する。図１１に各種文書論理要素の中でヘッダとフッタを不要要素として除去した結果を示す。

【００３８】なお、ヘッダ領域やフッタ領域であっても、巻・号・書誌情報などを含むものは１つにまとめて当該文書のメタデータとして利用できるため、例えば１ページ目の文書にあるヘッダ領域やフッタ領域については残しておくような構成にしても良い。どのページにヘッダ領域やフッタ領域を残しておくのかはオペレータが任意に指定できるものとする。

【００３９】また、予めオペレータが任意の文書論理要素を不要要素として指定しておくことで、その指定された文書論理要素を選択的に除去するような構成であっても良い。つまり、ヘッダやフッダを不要要素として予め指定しておけば、これを除去した形で構造化文書を得ることができる。

【００４０】文書論理要素グループ化部２２では、読み順情報が連続する文書論理要素のうち、

・図と図キャプション（図標題）と図脚注の組
・表と表キャプション（表標題）と表脚注の組
・表の中の連続するセルの組
・同じヘッディング記述を持つ箇条書きの組
・章節タイトルから次に出現する章節タイトルの直前の文書論理要素までの組

などを抽出し、それぞれを図グループ、表グループ、表カラム、箇条書きグループ、章節グループといったように１つにまとめてグループ化する。図１２にグループ化

の例を示す。

【００４１】図グループと表グループに関しては、それぞれに隣接する２つの文書論理要素領域Ｏｉ、Ｏｊが同一ページ上に存在し、以下に示す条件式を満たしている場合にグループ化する。これは、図や表は同一ページ内にまとめて割り付けられていることが多いという経験的な法則に基づいた条件である。

【００４２】キャプションが横書きの場合

・ｍｉｎ（ＯｉＸ２，ＯｊＸ２）－ｍａｘ（ＯｉＸ１，ＯｊＸ１）＞ｔｈ１、かつ、ｍａｘ（ＯｉＹ１，ＯｊＹ１）－ｍｉｎ（ＯｉＹ２，ＯｊＹ２）＜ｔｈ２

・キャプションが縦書きの場合

ｍｉｎ（ＯｉＹ２，ＯｊＹ２）－ｍａｘ（ＯｉＹ１，ＯｊＹ１）＞ｔｈ１、かつ、ｍａｘ（ＯｉＸ１，ＯｊＸ１）－ｍｉｎ（ＯｉＸ２，ＯｊＸ２）＜ｔｈ２

ただし、

ＯＸ１：文書論理要素領域Ｏの外接矩形の左上端Ｘ座標、ＯＸ２：文書論理要素領域Ｏの外接矩形の右下端Ｘ座標、ＯＹ１：文書論理要素領域Ｏの外接矩形の左上端Ｙ座標、ＯＹ２：文書論理要素領域Ｏの外接矩形の右下端Ｙ座標、ｔｈ１：閾値、ｔｈ２：閾値であり、文書画像の左上端を原点（０，０）とする。また、例えばｍｉｎ（ＯｉＸ２，ＯｊＸ２）とはＸ座標値ＯｉＸ２とＯｊＸ２のうちの小さい方をとり、ｍａｘ（ＯｉＸ１，ＯｊＸ１）とはＸ座標値ＯｉＸ１とＯｊＸ１のうちの大きい方をとることを意味する。

【００４３】文書構造抽出部２３では、文書論理要素グループ化部２２によってグループ化された文書論理要素に基づいて、図１３に示すような木構造により記述された文書階層構造を抽出する。この場合、全体文書というルートノードがあり、その下に４つの章節グループが同階層に並び、各章節グループの下の階層に図グループや表グループや箇条書きグループが配置され、その下の階層に各グループを構成する論理要素領域が位置付けられるようになっている。つまり、図グループ、表グループ、表グループ、箇条書きグループ、章節グループはそれぞれ図部分木、表部分木、箇条書き部分木、章節部分木として全体木構造を構成する部分木として記述される。さらに、表や図形や写真／絵などのレイアウト要素や各文書論理要素は全体木構造のノードとして記述される。

【００４４】また、１つの章節部分木の中には、図１３のように、図部分木（図グループ）や表部分木（表グループ）や箇条書き部分木（箇条書きグループ）などが混在することがある。この場合、文書の印刷時に紙面上のスペースの都合（組版上の都合）によって、図や表、脚注などの浮遊要素が意図的に本文中に割り付けられるケースが存在することを考慮すると、これらの浮遊要素が木構造において必ずしも適切な箇所に配置されているとは限らない。

【００４５】そこで、文書構造変更部２４では、このようなケースを見つけ出し、文書構造抽出部２３で得られた木構造に対して図部分木や表部分木などの配置を変更する。ここで、図部分木や表部分木の変更先として、当該図あるいは表が存在する章節グループの末尾を対象とした場合には、図１４のように木構造が変更される。図１４（ａ）は変更前の木構造、同図（ｂ）は変更後の木構造であり、図部分木と表部分木が章節グループの末尾に変更された場合が示されている（図中の点線枠参照）。

【００４６】また、図部分木や表部分木の変更先として、文書の末尾を対象とする場合には図１５のように木構造が変更される。図１５（ａ）は変更前の木構造、同図（ｂ）は変更後の木構造であり、図部分木と表部分木が文書の末尾に変更された場合が示されている（図中の点線枠参照）。

【００４７】また、図や表を最初に参照している文章の文書論理要素領域あるいは最も多く参照している文章の文書論理要素領域を最適位置として検出し、その領域の直後に図部分木や表部分木を挿入することでも良い。

【００４８】脚注などについても同様に考えることができる。

【００４９】また、図や表などがグループではなく、キャプション（標題）などを伴わずに単独で出現する場合には、その図や表の単独の領域で変更処理を行う。また、このように単独で出現する場合に、その変更処理を実施しないようにしても良い。あるいは、図や表の前後の文書論理要素を統合した後に、その直後に図や表を移動させるようにしても良い。

【００５０】このように、文書構造において、図部分木（図グループ）や表部分木（表グループ）の変更箇所を幾通りか考えることができる。そこで、変更可能な箇所の候補（章節末尾、文書末尾、参照箇所など）をオペレータに提示することで、オペレータの指定に従って文書構造の領域配置を図１４（ｂ）や図１５（ｂ）のように変更するものとする。

【００５１】ところで、図や表はカラム間やページ間などに跨って紙面上に割り付けられている場合が多い。このような場合、元々１つの文書論理要素であったものが図や表の介在により複数の文書論理要素に分割された状態にある。そこで、文書論理要素統合部２５では、文書構造変更部２４により図や表の配置が変更されたときに、その前後の文書論理要素を統合して１つの文書論理要素にまとめる処理を行う。

【００５２】図１４（ｃ）や図１５（ｃ）に２つの段組領域（段組の文書論理要素）を１つに統合した状態を示す。この２つの段組領域は、図１４（ａ）や図１５（ａ）に示すように表グループの前後に存在していたものである。ここで、表グループの配置変更に伴い、図１４（ｂ）や図１５（ｂ）のように上記２つの段組領域が

隣接するようになったので、統合処理にて１つの段組領域としてまとめる処理を行う。言い換えれば、文章間に挿入されていた表を他の場所に移動させたことで、その表の前後にあった２つの文章を１つの連続した文章として統合するといった処理を施すものである。これは、表に限らず、図の配置を変更した場合などでも同様である。

【００５３】このときの統合処理は以下のような条件に基づいて行われる。

【００５４】すなわち、今、例えば段組などの文書論理要素領域Ｏｉと文書論理要素領域Ｏｊがあり、図１６（ａ）に示すように、それぞれ範囲が矩形領域で示されるものとする。文書論理要素領域Ｏｉの位置座標は（ＯｉＸ１，ＯｉＹ１）と（ＯｉＸ２，ＯｉＹ２）、文書論理要素領域Ｏｊの位置座標は（ＯｊＸ１，ＯｊＹ１）と（ＯｊＸ２，ＯｊＹ２）である。

【００５５】ここで、ＯｉとＯｊの２つの文書論理要素が同一ページのＩＤ情報と同じ論理属性を有すると共にその読み順が文書構造変更後に連続している場合において、以下のような条件式を満たすときに、カラムを跨った連続性を有する文書論理要素であると見なして、同図（ｂ）に示すように１つの文書論理要素として統合する。

【００５６】・横書きの場合

ｍｉｎ（ＯｉＸ２，ＯｊＸ２）－ｍａｘ（ＯｉＸ１，ＯｊＸ１）＋１≦０

・縦書きの場合

ｍｉｎ（ＯｉＹ２，ＯｊＹ２）－ｍａｘ（ＯｉＹ１，ＯｊＹ１）＋１≦０

なお、上記条件式を満たすことの他に、（１）前方の文書論理要素領域Ｏｉにおいて、文章の末尾行にハードリターンがなく、（２）後方の文書論理要素Ｏｊにおいて、文章の先頭行が字下げされてない（つまり先頭位置に文字が存在する）といったことを条件に加えることで、統合処理をより精度良く行うことができる。

【００５７】また、ＯｉとＯｊが同一の論理属性を有し、その読み順が文書構造変更後に連続しており、ページＩＤ情報が異なる場合には、上記条件式に基づいて、この２つの文書論理要素はページを跨った文書論理要素であると見なして、１つの文書論理要素として統合することでも良い。

【００５８】このような統合処理の後、文書出力部１７では、最終的に得られた木構造を元にして、図１７のように各要素毎にタグ付けされたＸＭＬ文書やＨＴＭＬ文書などの構造化文書を生成して出力する。なお、構造化文書を生成する手法については公知であるため、ここではその説明は省略するものとする。

【００５９】このように、複数ページからなる印刷文書を処理対象とし、そこに含まれる文章とは別の図、表などの浮遊要素を適切な位置にオペレータに指示などに従って設定した形でＸＭＬ文書やＨＴＭＬ文書などのタグ付けされた構造化文書を生成することができる。これにより、多様なアプリケーションへの適用性の高い構造化文書を容易に得ることができる。

【００６０】なお、上記実施形態では、浮遊要素として、図と表を対象として説明したが、図と表の他に例えば写真、絵、脚注などの文書論理要素についても浮遊要素として同様に扱うことができるものである。

【００６１】また、上述した実施形態において記載した手法は、コンピュータに実行させることのできるプログラムとして、例えば磁気ディスク（フレキシブルディスク、ハードディスク等）、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤ等）、半導体メモリなどの記録媒体に書き込んで各種装置に適用したり、そのプログラム自体をネットワーク等の伝送媒体により伝送して各種装置に適用することも可能である。本装置を実現するコンピュータは、記録媒体に記録されたプログラムあるいは伝送媒体を介して提供されたプログラムを読み込み、このプログラムによって動作が制御されることにより、上述した処理を実行する。

【００６２】

【発明の効果】以上詳記したように本発明によれば、複数ページからなる印刷文書からＸＭＬ文書やＨＴＭＬ文書のようなタグ付けされた構造化文書を高精度に生成することができる。この場合、文章とは別の図や表などの浮遊要素を抽出し、生成される構造化文書の適切な箇所あるいはオペレータによって指定された箇所に設定することができるので、大量文書の電子化においてオペレータによる編集作業を大幅に軽減することができるものである。

【図面の簡単な説明】

【図１】本発明の処理対象を説明するための図であり、図１（ａ）は処理対象となる印刷文書に対応した文書画像の一例、同図（ｂ）はその文書画像に対応した構造化文書の一例を示す図。

【図２】本発明の一実施形態に係る文書処理装置の機能構成を示すブロック図。

【図３】上記文書処理装置のレイアウト解析部による処理結果を示す図であって、図３（ａ）は印刷文書に対応した文書画像、同図（ｂ）はその文書画像の部分領域（レイアウト要素群）を示す図。

【図４】レイアウト解析結果として得られる文章領域の構成を示す図。

【図５】上記文章領域の位置座標を示す図。

【図６】上記文章領域の木構造（階層構造）を示す図。

【図７】上記文書処理装置の文書論理要素抽出部による処理結果を示す図であって、図７（ａ）は文書画像の部分領域（レイアウト要素群）、同図（ｂ）はその部分領域から抽出された文書論理要素領域を示す図。

【図８】上記文書処理装置の読み順設定部による処理結

果を示す図であって、図８（ａ）はページ単位で読み順を付した場合、同図（ｂ）はページの連続性に伴って読み順を付した場合を示す図。

【図９】上記文書処理装置のレイアウト解析部、文字認識部、文書論理要素抽出部、読み順設定部による処理結果を示す図であって、図９（ａ）は文書画像の文書論理要素領域とその読み順を示す図、同図（ｂ）はその文書論理要素領域の木構造を示す図、同図（ｃ）はその文書論理要素領域のデータ構成を示す図。

【図１０】上記文書処理装置の文書構造解析部の構成を示すブロック図。

【図１１】上記文書処理装置の不要文書論理要素除去部によって文書論理要素領域から不要要素（ヘッダとフッタ）を除去した結果を示す図。

【図１２】上記文書処理装置の文書論理要素グループ化部による処理結果を示す図。

【図１３】上記文書処理装置の文書構造抽出部による処理結果を示す図。

【図１４】上記文書処理装置の文書構造変更部と文書論理要素統合部による処理結果（その１）を示す図であって、図１４（ａ）は変更前の木構造、同図（ｂ）は変更後の木構造、同図（ｃ）は文書論理要素統合後の木構造を示す図。

【図１５】上記文書処理装置の文書構造変更部と文書論理要素統合部による処理結果（その２）を示す図であって、図１５（ａ）は変更前の木構造、同図（ｂ）は変更後の木構造、同図（ｃ）は文書論理要素統合後の木構造を示す図。

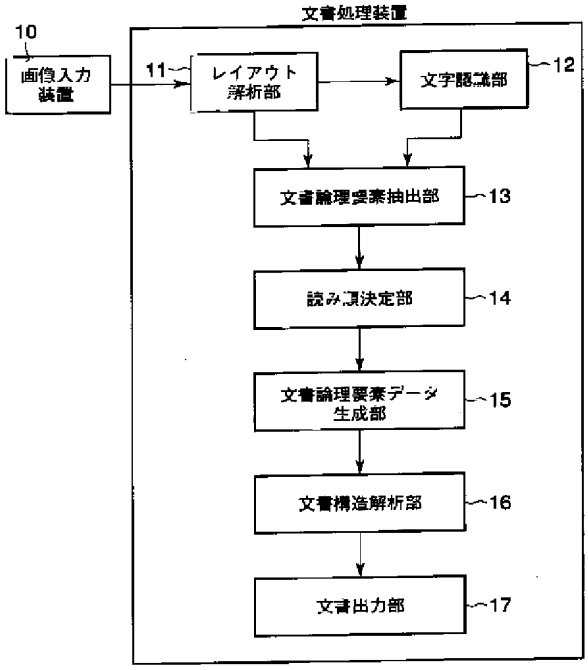
【図１６】上記文書処理装置の文書論理要素統合部による統合処理を説明するための図であって、図１６（ａ）は統合前の２つの文書論理要素の位置座標を示す図、同図（ｂ）は２つの文書論理要素を統合した場合を示す図。

【図１７】上記文書処理装置の文書出力部による処理結果（構造化文書）を示す図。

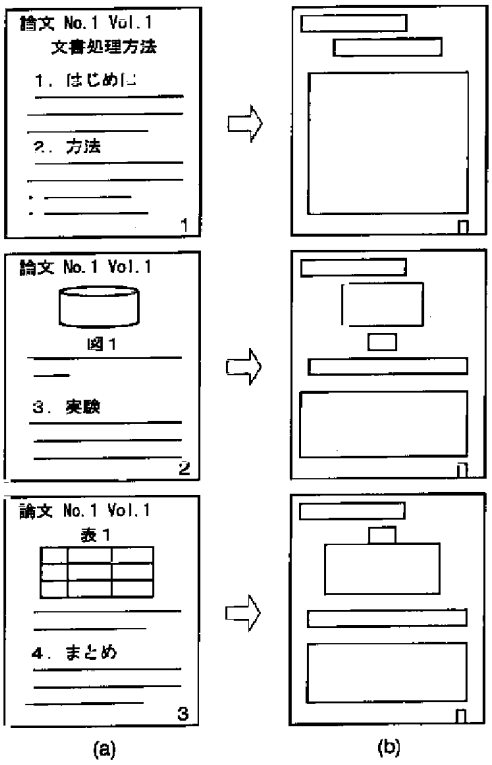
【符号の説明】

１０…文書入力装置
１１…レイアウト解析部
１２…文字認識部
１３…文書論理要素抽出部
１４…読み順設定部
１５…文書論理要素データ生成部
１６…文書構造解析部
１７…文書出力部
２１…不要文書論理要素除去部
２２…文書論理要素グループ化部
２３…文書構造抽出部
２４…文書構造変更部
２５…文書論理要素統合部

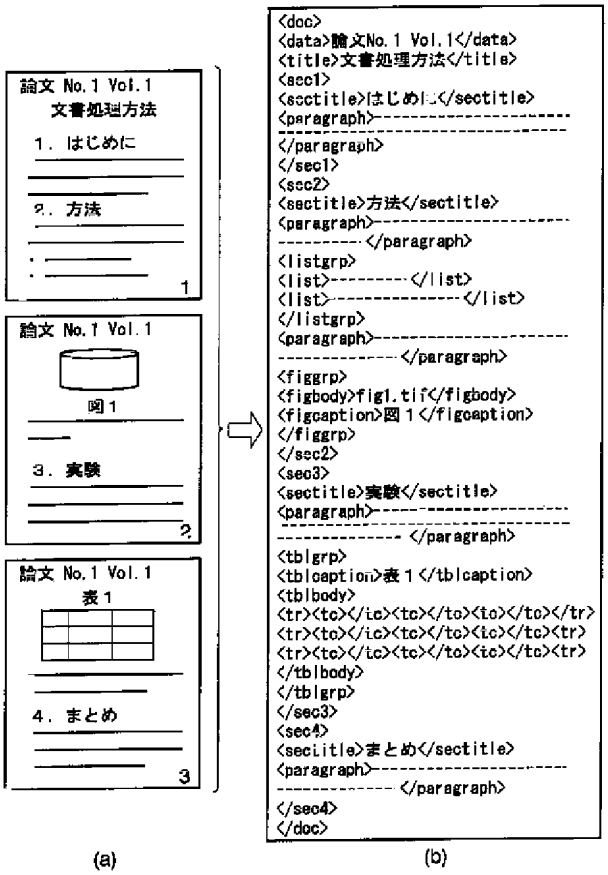
【図２】



【図３】

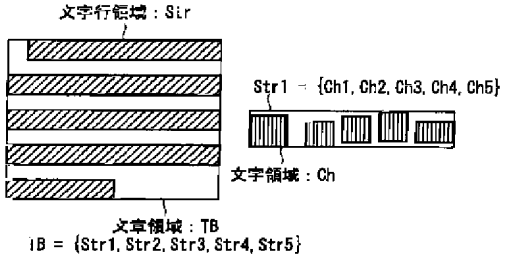


【図1】

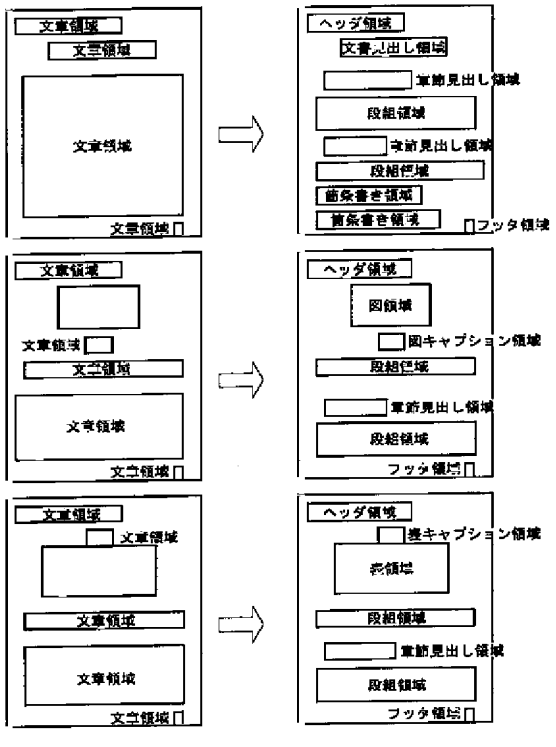

論文 No.1 Vol.1
文書処理方法
1．はじめに
2．方法
1
論文 No.1 Vol.1
図1
3．実験
2
論文 No.1 Vol.1
表1
4．まとめ
3
(a)
<doc>
<data>論文No.1 Vol.1</data>
<title>文書処理方法</title>
<sec1>
<sectitle>はじめに</sectitle>
<paragraph>----------
</paragraph>
</sec1>
<sec2>
<sectitle>方法</sectitle>
<paragraph>----------
----------</paragraph>
<listgrp>
<list>----------</list>
<list>----------</list>
</listgrp>
<paragraph>----------
----------</paragraph>
<figgrp>
<figbody>fig1.tif</figbody>
<figcaption>図1</figcaption>
</figgrp>
</sec2>
<sec3>
<sectitle>実験</sectitle>
<paragraph>----------

---------- </paragraph>
<tblgrp>
<tblcaption>表1</tblcaption>
<tblbody>
<tr><tc></tc><tc></tc><tc></tc></tr>
<tr><tc></tc><tc></tc><tc></tc><tr>
<tr><tc></tc><tc></tc><tc></tc><tr>
</tblbody>
</tblgrp>
</sec3>
<sec4>
<sectitle>まとめ</sectitle>
<paragraph>----------
---------- </paragraph>
</sec4>
</doc>
(b)

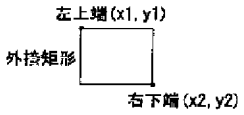
【図4】


文字行領域：Str
Str1 = {Ch1, Ch2, Ch3, Ch4, Ch5}
文字領域：Ch
文章領域：TB
TB = {Str1, Str2, Str3, Str4, Str5}

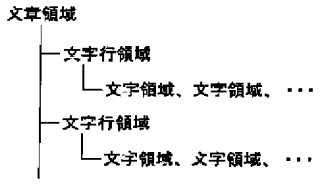
【図7】


文章領域
文字領域
文章領域
文字領域
ヘッダ領域
文書見出し領域
章節見出し領域
段組領域
節見出し領域
段組領域
箇条書き領域
箇条書き領域
フッタ領域
図領域
図キャプション領域
表領域
表キャプション領域
(a)
(b)

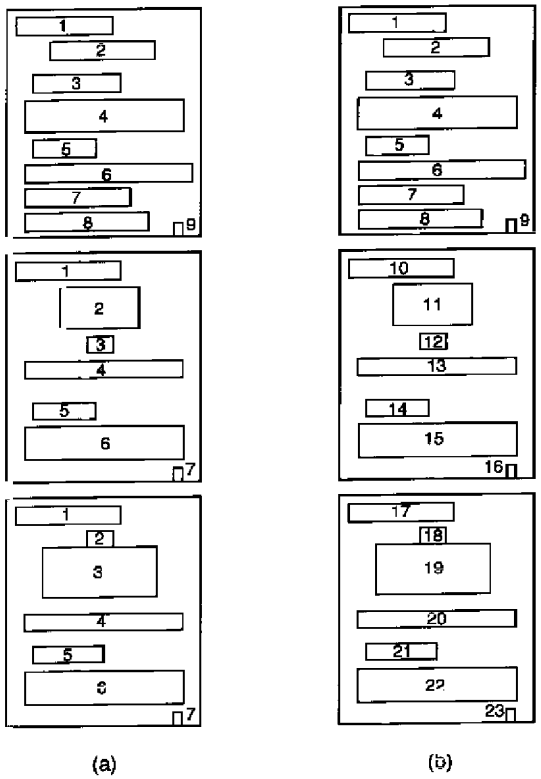
【図5】


左上端(x1, y1)
外接矩形
右下端(x2, y2)

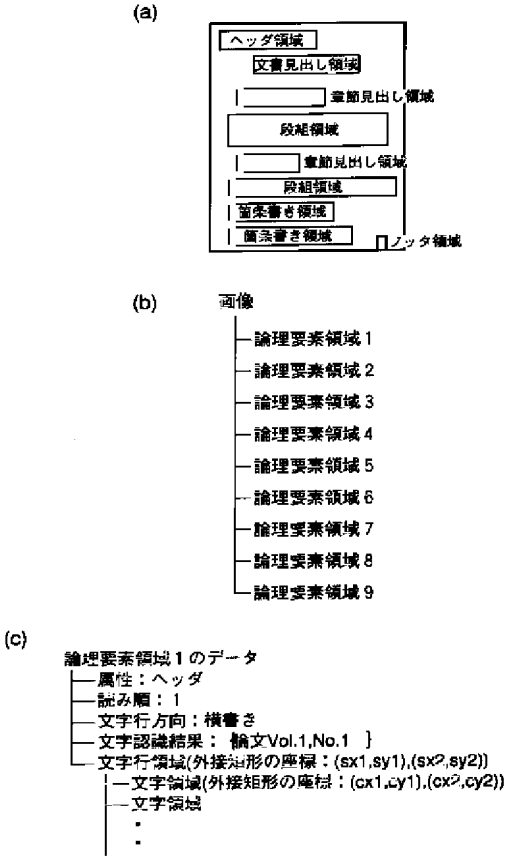
【図6】


文章領域
文字行領域
文字領域、文字領域、・・・
文字行領域
文字領域、文字領域、・・・

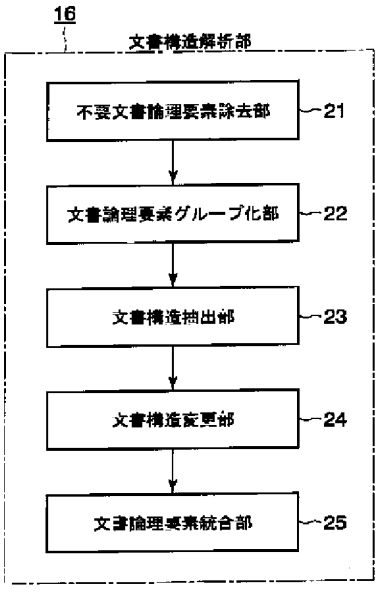
【図８】


(a)
(b)

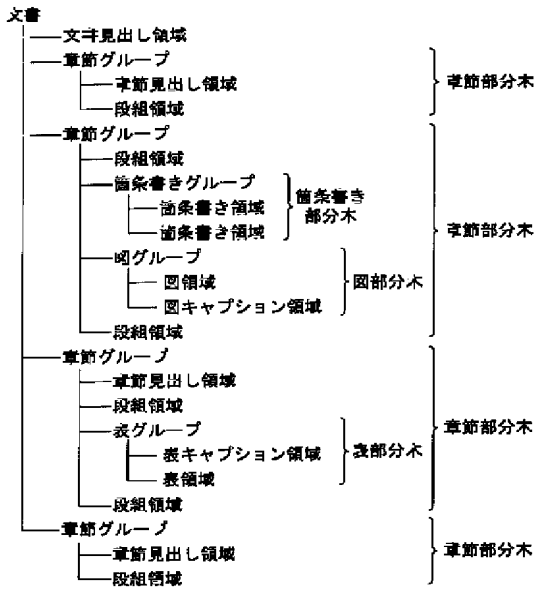
【図９】


(a)
ヘッダ領域
文書見出し領域
章節見出し領域
段組領域
章節見出し領域
段組領域
箇条書き領域
箇条書き領域
フッタ領域
(b)
画像
論理要素領域１
論理要素領域２
論理要素領域３
論理要素領域４
論理要素領域５
論理要素領域６
論理要素領域７
論理要素領域８
論理要素領域９
(c)
論理要素領域１のデータ
属性：ヘッダ
読み順：１
文字行方向：横書き
文字認識結果：　論文Vol.1,No.1　}
文字行領域(外接矩形の座標：(sx1,sy1),(sx2,sy2))
文字領域(外接矩形の座標：(cx1,cy1),(cx2,cy2))
文字領域

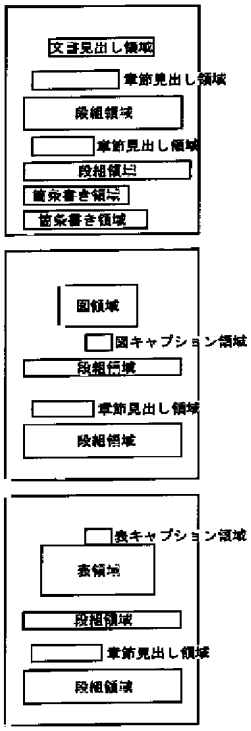
【図１０】


16
文書構造解析部
不要文書論理要素除去部
21
文書論理要素グループ化部
22
文書構造抽出部
23
文書構造変更部
24
文書論理要素統合部
25

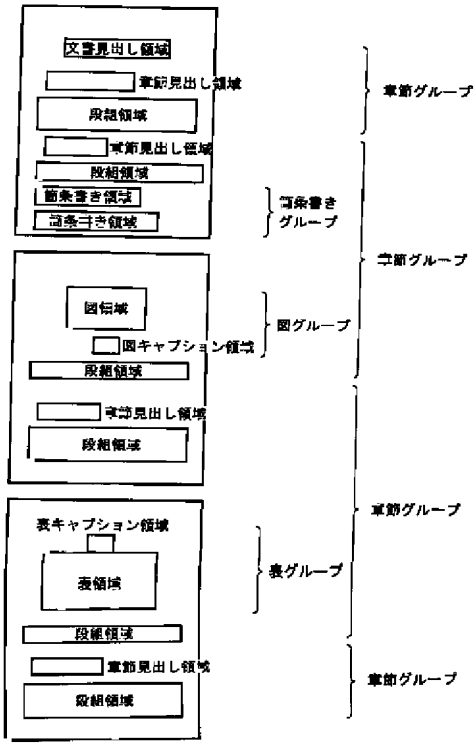
【図１３】


文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節部分木
章節グループ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
箇条書き部分木
図グループ
図領域
図キャプション領域
図部分木
段組領域
章節部分木
章節グループ
章節見出し領域
段組領域
表グループ
表キャプション領域
表領域
表部分木
段組領域
章節部分木
章節グループ
章節見出し領域
段組領域
章節部分木

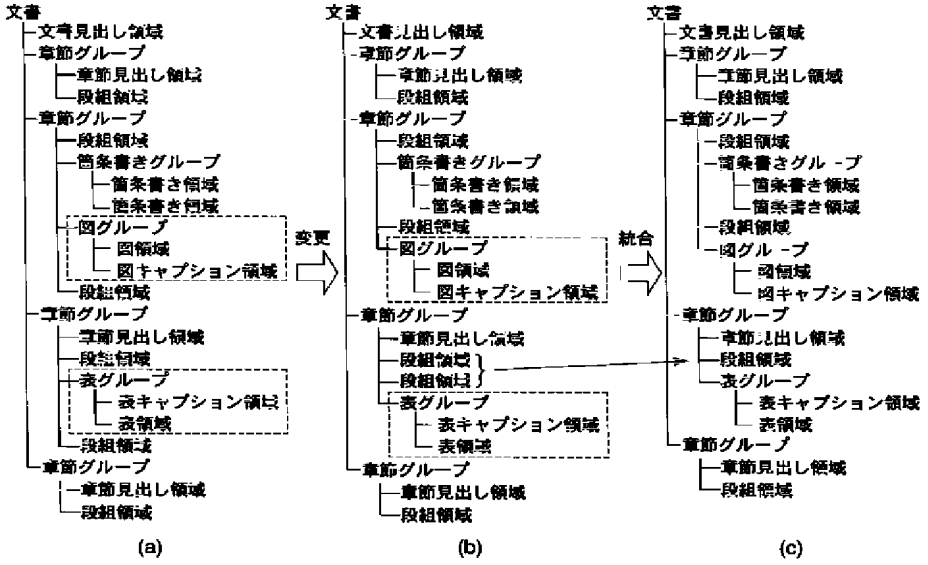
【図１１】


文書見出し領域
章節見出し領域
段組領域
章節見出し領域
段組領域
箇条書き領域
箇条書き領域
図領域
図キャプション領域
段組領域
章節見出し領域
段組領域
表キャプション領域
表領域
段組領域
章節見出し領域
段組領域

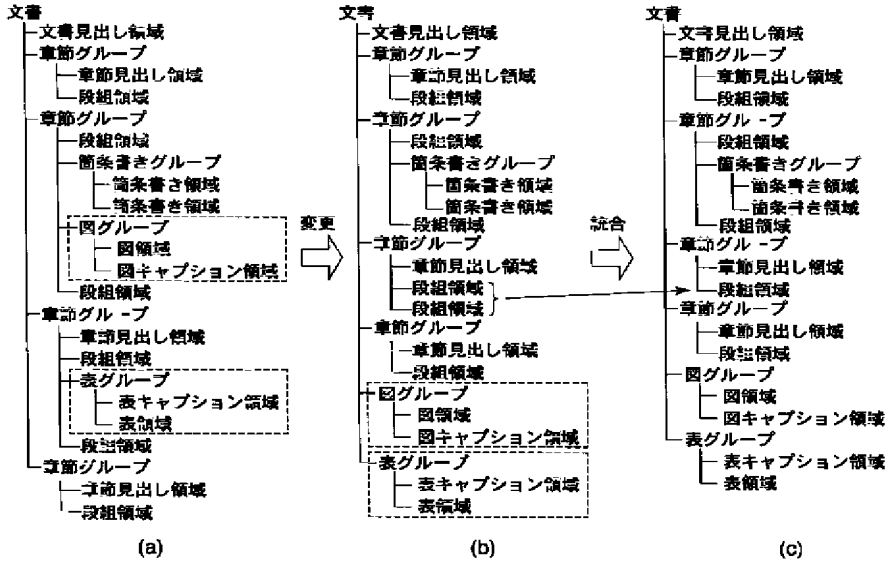
【図１２】


文書見出し領域
章節見出し領域
段組領域
章節見出し領域
段組領域
箇条書き領域
箇条書き領域
箇条書きグループ
章節グループ
図領域
図キャプション領域
図グループ
段組領域
章節見出し領域
段組領域
章節グループ
表キャプション領域
表領域
表グループ
段組領域
章節見出し領域
段組領域
章節グループ

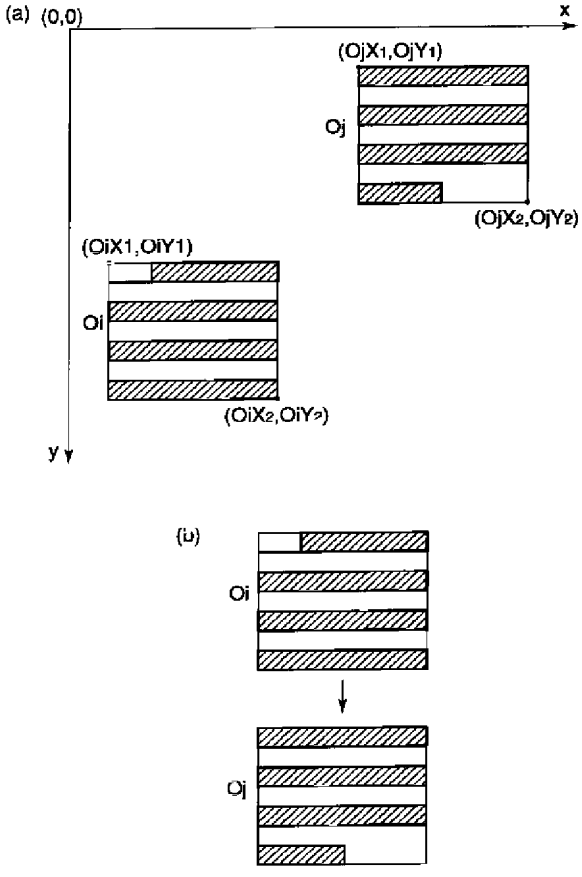
【図１４】


文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節グループ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
図グループ
図領域
図キャプション領域
段組領域
章節グループ
章節見出し領域
段組領域
表グループ
表キャプション領域
表領域
段組領域
章節グループ
章節見出し領域
段組領域
変更
文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節グループ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
段組領域
図グループ
図領域
図キャプション領域
章節グループ
章節見出し領域
段組領域
段組領域
表グループ
表キャプション領域
表領域
章節グループ
章節見出し領域
段組領域
統合
文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節グループ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
段組領域
図グループ
図領域
図キャプション領域
章節グループ
段組領域
表グループ
表キャプション領域
表領域
章節グループ
章節見出し領域
段組領域
(a)
(b)
(c)

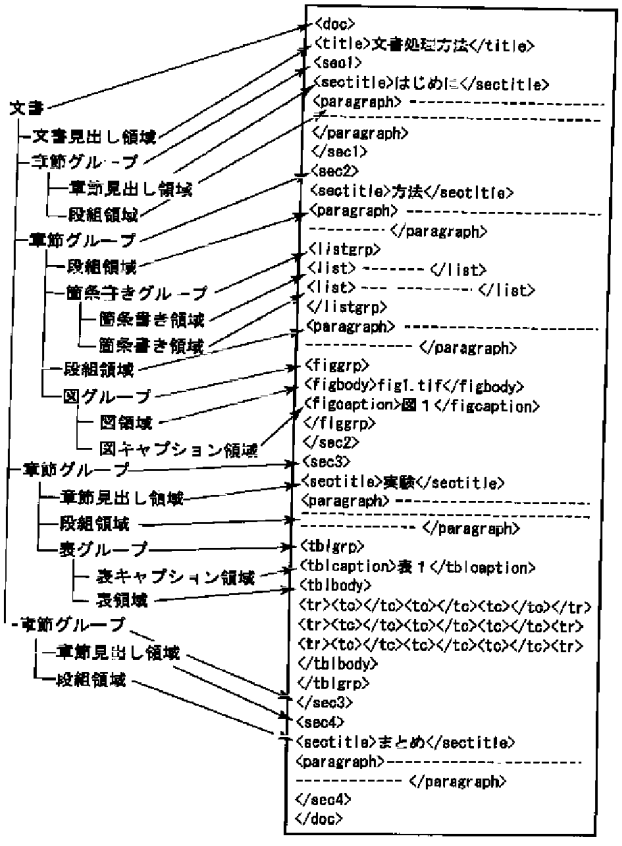
【図15】


文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節グループ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
図グループ
図領域
図キャプション領域
段組領域
章節グル -プ
章節見出し領域
段組領域
表グループ
表キャプション領域
表領域
段組領域
章節グループ
章節見出し領域
段組領域
変更
(a)
文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節グループ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
段組領域
章節グループ
章節見出し領域
段組領域
段組領域
章節グループ
章節見出し領域
段組領域
図グループ
図領域
図キャプション領域
表グループ
表キャプション領域
表領域
(b)
統合
文書
文書見出し領域
章節グループ
章節見出し領域
段組領域
章節グル -プ
段組領域
箇条書きグループ
箇条書き領域
箇条書き領域
段組領域
章節グル -プ
章節見出し領域
段組領域
章節グループ
章節見出し領域
段組領域
図グループ
図領域
図キャプション領域
表グループ
表キャプション領域
表領域
(c)

【図16】


(a)
(0,0)
x
(OjX1,OjY1)
Oj
(OjX2,OjY2)
(OiX1,OiY1)
Oi
(OiX2,OiY2)
y
(b)
Oi
Oj

【図17】


文書
文書見出し領域
章節グル -プ
章節見出し領域
段組領域
章節グループ
段組領域
箇条書きグル -プ
箇条書き領域
箇条書き領域
段組領域
図グループ
図領域
図キャプション領域
章節グループ
章節見出し領域
段組領域
表グループ
表キャプション領域
表領域
章節グループ
章節見出し領域
段組領域
<doc>
<title>文書処理方法</title>
<sec1>
<sectitle>はじめに</sectitle>
<paragraph> ----------
</paragraph>
</sec1>
<sec2>
<sectitle>方法</sectitle>
<paragraph> ----------
---------- </paragraph>
<listgrp>
<list> ------- </list>
<list> --- --------- </list>
</listgrp>
<paragraph> ----------
---------- </paragraph>
<figgrp>
<figbody>fig1.tif</figbody>
<figcaption>図１</figcaption>
</figgrp>
</sec2>
<sec3>
<sectitle>実験</sectitle>
<paragraph> ----------
---------- </paragraph>
<tblgrp>
<tblcaption>表１</tblcaption>
<tblbody>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</tblbody>
</tblgrp>
</sec3>
<sec4>
<sectitle>まとめ</sectitle>
<paragraph>----------
---------- </paragraph>
</sec4>
</doc>